



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Complete genome sequences of elephant endotheliotropic herpesviruses 1A and 1B determined directly from fatal cases

Citation for published version:

Wilkie, GS, Davison, AJ, Watson, M, Kerr, K, Sanderson, S, Bouts, T, Steinbach, F & Dastjerdi, A 2013, 'Complete genome sequences of elephant endotheliotropic herpesviruses 1A and 1B determined directly from fatal cases', *Journal of Virology*, vol. 87, no. 12, pp. 6700-6712. <https://doi.org/10.1128/JVI.00655-13>

Digital Object Identifier (DOI):

[10.1128/JVI.00655-13](https://doi.org/10.1128/JVI.00655-13)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

Journal of Virology

Publisher Rights Statement:

Freely available in PMC.

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Complete Genome Sequences of Elephant Endotheliotropic Herpesviruses 1A and 1B Determined Directly from Fatal Cases

Gavin S. Wilkie,^a Andrew J. Davison,^a Mick Watson,^b Karen Kerr,^a Stephanie Sanderson,^c Tim Bouts,^{d*} Falko Steinbach,^e Akbar Dastjerdi^e

MRC—University of Glasgow Centre for Virus Research, Glasgow, United Kingdom^a; ARK-Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, United Kingdom^b; Chester Zoo, Chester, United Kingdom^c; ZSL Whipsnade Zoo, Dunstable, United Kingdom^d; Virology Department, Animal Health and Veterinary Laboratories Agency Weybridge, Addlestone, United Kingdom^e

A highly lethal hemorrhagic disease associated with infection by elephant endotheliotropic herpesvirus (EEHV) poses a severe threat to Asian elephant husbandry. We have used high-throughput methods to sequence the genomes of the two genotypes that are involved in most fatalities, namely, EEHV1A and EEHV1B (species *Elephantid herpesvirus 1*, genus *Proboscivirus*, subfamily *Betaherpesvirinae*, family *Herpesviridae*). The sequences were determined from postmortem tissue samples, despite the data containing tiny proportions of viral reads among reads from a host for which the genome sequence was not available. The EEHV1A genome is 180,421 bp in size and consists of a unique sequence (174,601 bp) flanked by a terminal direct repeat (2,910 bp). The genome contains 116 predicted protein-coding genes, of which six are fragmented, and seven paralogous gene families are present. The EEHV1B genome is very similar to that of EEHV1A in structure, size, and gene layout. Half of the EEHV1A genes lack orthologs in other members of subfamily *Betaherpesvirinae*, such as human cytomegalovirus (genus *Cytomegalovirus*) and human herpesvirus 6A (genus *Roseolovirus*). Notable among these are 23 genes encoding type 3 membrane proteins containing seven transmembrane domains (the 7TM family) and seven genes encoding related type 2 membrane proteins (the EE50 family). The EE50 family appears to be under intense evolutionary selection, as it is highly diverged between the two genotypes, exhibits evidence of sequence duplications or deletions, and contains several fragmented genes. The availability of the genome sequences will facilitate future research on the epidemiology, pathogenesis, diagnosis, and treatment of EEHV-associated disease.

The Asian elephant (*Elephas maximus*) is listed by the International Union for Conservation of Nature as endangered (www.iucnredlist.org/details/7140/0). This iconic animal is declining in the wild largely because of loss, degradation, and fragmentation of its habitat. Moreover, elephant husbandry in zoos is facing the challenge of a hemorrhagic disease associated with infection by elephant endotheliotropic herpesvirus (EEHV) (1–3). This highly lethal disease was first reported in 1990 (4). Death typically occurs in young elephants (1 to 4 years old) after a short clinical period of up to a week, characterized by lethargy, anorexia, generalized edema of the head, neck, trunk, and forelimbs, oral ulcerations, and cyanosis of the tongue (5). The disease is reported to have resulted in a mortality of at least 80% among the approximately 80 known victims over the past 20 years and to have been a major killer of monitored, younger animals (3).

The virus involved in most cases of fatal disease is elephant endotheliotropic herpesvirus 1 (EEHV1). This virus exists as two genotypes, namely, EEHV1A, which was discovered in 1999 (6), and EEHV1B, which was reported a couple of years later (7). Characterization of limited DNA sequences for EEHV1A and extensive DNA sequences for EEHV1B resulted in the classification of EEHV1 into the new species *Elephantid herpesvirus 1*, as the founding member of the genus *Proboscivirus*, subfamily *Betaherpesvirinae*, family *Herpesviridae*, order *Herpesvirales* (6, 8–10). Members of subfamily *Betaherpesvirinae* are commonly referred to as betaherpesviruses and, in addition to EEHV1, include viruses of humans and other primates in the genera *Cytomegalovirus* (e.g., human cytomegalovirus; HCMV) and *Roseolovirus* (human herpesviruses 6A, 6B, and 7; HHV6A, HHV6B, and HHV7, respectively), rodent viruses in the genus *Muromegalovirus*, and several

other viruses of nonprimates (tupaia, bat, pig, and guinea pig) that have not yet been placed into genera. Five additional EEHV genotypes (EEHV2 to EEHV6) have been identified by limited genetic analyses (6, 11, 12), and data for a sixth (EEHV7) are available in NCBI GenBank (e.g., JQ300082). These genotypes are phylogenetically more distant from EEHV1A than is EEHV1B, but all are sufficiently closely related to EEHV1A to belong to the genus *Proboscivirus* (3).

Five of the eight EEHV genotypes (EEHV1A, EEHV1B, EEHV3, EEHV4, and EEHV5) have been reported in association with hemorrhagic disease in Asian elephants (6, 11–13). Also, five genotypes (EEHV1A, EEHV2, EEHV3, EEHV6, and EEHV7) have been detected in wild African elephants (*Loxodonta africana*) (3), but these appear to be generally less virulent in this host, with only EEHV2 having been associated with hemorrhagic disease (6). Indeed, herpesviruses were described in pulmonary nodules and cutaneous papillomas from African elephants long before they were discovered in Asian elephants (14, 15). Detection of EEHV1A in papillomas from African elephants led to the hypothesis that hemorrhagic disease in Asian elephants originated from interspecies

Received 6 March 2013 Accepted 26 March 2013

Published ahead of print 3 April 2013

Address correspondence to Andrew J. Davison, andrew.davison@glasgow.ac.uk.

* Present address: Tim Bouts, Al Wabra Wildlife Preservation, Al Shahaniya, State of Qatar.

Copyright © 2013, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JVI.00655-13

transfer events in situations where the two elephant species are housed in close proximity (6). Although transfer events of this kind may have occurred, subsequent investigations have indicated that the epidemiology of EEHV infections is more complex (3, 16, 17). For example, EEHVs may circulate asymptomatically in Asian elephant populations; thus, they do not necessarily rely on interspecies transfer for disease development (18–20).

The rapid, generalized damage to capillary endothelial cells that occurs in various organs during EEHV infection usually renders intervention practices futile. Thus, it is likely that EEHVs will continue to provoke losses among the already shrinking populations of Asian elephants. Antivirals directed against human herpesviruses have been administered in some cases, but it has not been established whether such treatments are effective or to what extent they have contributed to instances of survival (3, 6, 21). Furthermore, no vaccines are available yet. Basic research into characterizing viral growth properties, evaluating potential interventions, and producing serological reagents has been severely hampered by a continuing inability to grow EEHVs of any genotype in cell culture. As a result, determination of sequence data has depended on the laborious processes of PCR amplification from DNA isolated from postmortem tissues. In this article, we report the use of high-throughput methods to determine from postmortem tissues the complete genome sequences of strains of the two EEHV genotypes (EEHV1A and EEHV1B) that cause the majority of lethal disease in Asian elephants.

MATERIALS AND METHODS

Sources and typing of viral samples. Necropsy specimens were obtained from two juvenile Asian elephants that had been housed in separate zoos in the United Kingdom. In each case, death had been attributed to EEHV1 on the basis of clinical history and detection of viral DNA in various tissues by PCR. The elephants were Raman (male), who was born at Chester Zoo on 12 November 2006 and died on 23 July 2009 (http://www.elephant.se/database2.php?elephant_id=581), and Emelia (female; barn name; official name, Aneena), who was born at Whipsnade Zoo on 16 March 2004 and died on 17 December 2006 (http://www.elephant.se/database2.php?elephant_id=779). The viral genotypes were determined by PCR amplification and sequencing of a short region of the DNA polymerase gene (U38) and were assigned to EEHV1A for Raman and EEHV1B for Emelia (data not shown). The viruses were named EEHV1A strain Raman and EEHV1B strain Emelia.

Extraction of DNA from tissue samples. DNA was purified from frozen host tissues (heart for EEHV1A and tongue for EEHV1B) by using a DNeasy Blood & Tissue kit (Qiagen, Crawley, United Kingdom). Approximately 25 mg of tissue was homogenized in 360 μ l of kit buffer ATL by using a Mikro-Dismembrator S (Sartorius Mechatronics, Epsom, United Kingdom) at 3,000 rpm for 1 min. The lysate was centrifuged briefly, and 180 μ l of supernatant was mixed with 20 μ l of proteinase K from the kit and shaken at 5,000 rpm for 1 h at 56°C in a Thermomixer (Eppendorf, Stevenage, United Kingdom). DNA extraction was concluded according to the kit protocol. The EEHV1A and EEHV1B DNA concentrations were 86.8 and 93.2 ng/ μ l, respectively.

Generation, assembly, and verification of DNA sequence data. High-throughput sequence data sets were generated from the DNA samples in the form of 76-nucleotide (nt) paired-end reads by using a HiSeq 2000 instrument (Illumina, San Diego, CA) at ARK-Genomics (University of Edinburgh). DNA (5 μ g) from each sample was broken randomly by treatment with NEBNext double-stranded DNA (dsDNA) fragmentase (New England BioLabs, Ipswich, MA), and libraries for sequencing were constructed by using a TruSeq DNA sample preparation kit (Illumina). The read data were quality filtered, trimmed, processed for adapter removal, and subjected to read pair validation by using Trim Galore v. 0.2.2,

which is a wrapper for Cutadapt and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore).

To prepare data for *de novo* assembly, host reads were subtracted by using BWA v. 0.6.2-r126 (22) in gapped, paired-end mode to align trimmed reads to the African elephant genome (Ensembl Loxafr3.0 v. 69.3). Unmapped read pairs that potentially originated from viral DNA were then extracted by using Samtools v. 0.1.18 (23). Contigs were assembled from the quality-filtered, host-subtracted data sets in paired-end and single-end modes by using the overlap-layout-consensus assembly program Edena v. 3.120926 (24) and the de Bruijn graph assemblers Velvet v. 1.2.07 (25) and SSAKE v. 3.8 (26). Individual assembly algorithms tend to terminate contigs at different sites, producing contigs that overlap those produced by other algorithms. Therefore, the contigs obtained from each assembly were merged into larger contigs by using Phrap version 1.080812 (27, 28). Further joins were then made in this set of merged contigs by using the iterative assembly and gap elimination program IMAGE v. 2.31 (29), using the host-subtracted, paired-end read data to extend contigs and close gaps. All of the contigs from the various assemblies were then entered into a Gap4 database (30), and additional joins were made in regions of overlap. Final joins were made by using two iterative approaches. The first involved assembling data sets against contigs that had been extended speculatively (usually by employing mononucleotide tract sequences), relying on a degree of mismatch to permit incorporation of reads (31). The second approach involved using custom Perl scripts to extract from data sets individual reads that extended contigs (31, 32). The assembly programs used were BWA v. 0.6.2-r126 in paired-end mode and Maq v. 0.7.1 (33) in single-end mode. Alignments were visualized by using Tablet 1.12.12.05 (34).

The sequences of specific regions in the viral genomes were confirmed by direct Sanger sequencing of PCR products using standard techniques. The targeted regions included tandem repeats at nt 401 to 766, 1173 to 1306, 1570 to 1728, 111219 to 111378, 132916 to 132954, 167486 to 167518, 168960 to 169002, 169808 to 169859, 173671 to 173725, 177912 to 178277, 178684 to 178817, and 179081 to 179239 in the finished EEHV1A sequence and nt 400 to 731, 1138 to 1267, 1532 to 1703, 110799 to 110967, 132429 to 132475, 167054 to 167080, 168537 to 168563, 173595 to 173639, 177859 to 178190, 178597 to 178726, and 178991 to 179162 in the finished EEHV1B sequence. In addition, regions at the junctions between the terminal direct repeats (TR) and the unique sequence (U) were verified (nt 2839 to 3217 and 177416 to 177641 in the finished EEHV1A sequence).

Identification of the genome termini. The 5' termini of the EEHV1A genome were identified as described previously for other herpesviruses (35). This involved (i) ligating a partially double-stranded DNA adaptor to an aliquot of DNA that had been blunt ended, (ii) carrying out a first round of PCR using a primer (AP1) that matched part of the adaptor and another primer (GTL or GTR) that matched a sequence close to the putative left or right terminus, respectively, (iii) carrying out a second round of PCR on the products of the first round using a primer (AP2) that matched another part of the adaptor and GTL or GTR as appropriate (i.e., heminested PCR), and (iv) sequencing plasmid clones made from the PCR products generated. The adaptor oligonucleotides were 5'-CTA ATA CGA CTC ACT ATA GGG CTC GAG CGG CCG CCC GGG CAG GT-3' and 5'-ACC TGC CC-3'. The latter oligonucleotide was phosphorylated at its 5' end, blocked by an amino group at its 3' end and complementary to the 3' end of the longer oligonucleotide. AP1 was 5'-CCA TCC TAA TAC GAC TCA CTA TAG GGC-3', AP2 was ACT CAC TAT AGG GCT CGA GCG GCC GCC CGG GCA GGT-3', GTL was 5'-TCC GGG AGG ATA TAC GTC ACA ATG-3', and GTR was 5'-CTT CGT GAC GCG GCT CCC GTT AAT-3'.

The 3' termini of the EEHV1A genome were identified by a modification of the method described above, in which the longer adaptor oligonucleotide was replaced by 5'-CTA ATA CGA CTC ACT ATA GGG CTC GAG CGG CCG CCC GGG CAG GTN-3' (i.e., adding a single, redundant

nucleotide at the 3' end), the DNA was not blunt ended, and step iii was omitted because the second round of PCR proved to be unnecessary.

Nucleotide sequence accession numbers. The finished sequences of the EEHV1A and EEHV1B genomes were deposited in GenBank under accession numbers [KC462165](#) and [KC462164](#), respectively.

RESULTS AND DISCUSSION

Generation of genome sequences. The original, unfiltered read data sets for EEHV1A and EEHV1B consisted of 107,670,730 and 195,294,050 reads, respectively, all 76 nt in length. The data for each sample had been derived by using two lanes on the sequencing instrument; thus, they consisted of two approximately equal parts. *De novo* assembly was carried out by using one of these parts for each sample, first subtracting reads that were of low quality or that originated from host DNA. As the genome sequence of the Asian elephant was not available for the host subtraction step, that of the related African elephant was used instead. The quality-filtered, host-subtracted data sets consisted of 1,650,532 and 1,709,026 reads for EEHV1A and EEHV1B, respectively. These were assembled *de novo* using a range of alternative programs in order to profit from the advantages of each. Subsequent steps focused on EEHV1A, because the contigs obtained were consistently longer than those obtained for EEHV1B, regardless of the program used, and it was evident that the EEHV1B data provided lower coverage of viral contigs. Assembly of all the EEHV1A contigs generated by the various programs, followed by final joins made by iterative approaches, resulted in a circular contig of 178 kbp representing the viral genome, plus a large collection of much smaller contigs, which presumably originated from the Asian elephant genome. The smaller contigs had coverage depths that were very different from that of the viral contig, and many were repetitive in sequence.

Derivation of the EEHV1B sequence was predicated on the assumption that it would be similar to the EEHV1A sequence. The original, unfiltered EEHV1B data set was assembled against the EEHV1A sequence, necessary corrections were made, and iterative approaches were used to extend and join contigs. The resulting contigs were then supplemented and extended by the many contigs generated for EEHV1B by *de novo* assembly, and remaining gaps were closed by using iterative approaches. This produced a single, circular contig closely similar in size to that of EEHV1A.

Various types of tandem repeat, at which a degree of length variation was possible, were evident in the sequences of the EEHV1A and EEHV1B contigs. Each repeat was examined individually, and necessary adjustments were made to the sequence. The first type of repeat consisted of a mononucleotide (homopolymer) tract (G:C or A:T) of >8 bp. EEHV1A contained 10 G:C and 15 A:T tracts, and EEHV1B contained 11 G:C and 10 A:T tracts. The second type of repeat consisted of a dinucleotide tract [(AC):(GT)] of >20 bp. EEHV1A contained six such repeats, and EEHV1B contained five. The mode length of each mono- and dinucleotide repeat was examined by extracting the relevant, individual reads from the original, unfiltered data sets. The lengths of five of the dinucleotide repeats in EEHV1A and four in EEHV1B were also assessed by PCR and sequencing and were fully concordant with those derived by examination of reads. The third type of repeat consisted of sequences in which, unlike the mono- and dinucleotide repeats, the length of the repeat exceeded that of the read length (76 nt), rendering the sequences unresolvable from the high-throughput data alone. EEHV1A and EEHV1B

each contained four such repeats, the sequences and lengths of which were determined by PCR and sequencing.

The final stage in determining the genome sequences involved identifying the termini. Since all sequenced herpesvirus genomes are linear (10), the derivation of circular contigs for EEHV1A and EEHV1B implied that each genome contains a terminal direct repeat longer than the read length. Bearing this in mind, candidate termini for EEHV1A were identified by searching manually for regions that are similar in sequence to the termini of other beta-herpesvirus genomes and located close to unusually large sets of identical reads that may have originated from the termini, on the basis that all other reads should be spread randomly throughout the genome. The strongest predictions for the left and right termini were supported by 12 and 13 reads, respectively, and were investigated experimentally by a PCR-based method (see Materials and Methods). The PCR products generated were 190 and 131 bp in size for the left and right terminus, respectively (data not shown). Sequencing of clones of these PCR products located each 5' terminus to a specific nucleotide (supported by 11/12 and 23/24 clones for the left and right terminus, respectively). Since the herpesvirus genomes that have been investigated in sufficient detail contain an unpaired, complementary nucleotide at each 3' terminus, this possibility was investigated experimentally by a PCR-based method (see Materials and Methods). Sequencing of clones of the PCR products (data not shown) identified an unpaired, complementary C residue at the right 3' terminus and an unpaired, complementary G residue at the left 3' terminus (supported by 8/10 and 9/10 clones, respectively). This arrangement produces an added element of symmetry in the genome, in that the sequence generated from a conceptual fusion of the termini (such as might be present in replicating viral DNA) is identical to that at the U-TR junction, with a 16-bp region at the right end of U (i.e., outside TR) being repeated at the right terminus (i.e., inside TR) (Fig. 1a).

The locations of the EEHV1B genome termini were inferred from their locations in EEHV1A and are positioned in regions in which the two genomes are closely similar in sequence. It should be noted that although the locations of genome termini were predicted and then confirmed experimentally for EEHV1A, the directed nature of the search cannot rule out the existence of alternative termini in some molecules, similar to those reported for guinea pig cytomegalovirus (36) and HCMV (37), which yield genomes shorter than the standard forms. The complete EEHV1A and EEHV1B genome sequences were reconstructed from the sequences of the circular contigs on the basis of the locations of the termini. These finished sequences were checked for accuracy by aligning them against the original, unfiltered data sets and for continuity by ensuring that the coverage of read pairs was uninterrupted throughout.

The EEHV1A and EEHV1B genomes consist of a large, unique sequence flanked by a terminal direct repeat, in the arrangement TR-U-TR. This structure is common to several characterized beta-herpesvirus genomes, including those of HHV6A, HHV6B, and HHV7 (10). The orientation of the EEHV1A and EEHV1B genome sequences was assigned on the basis of similarities between the sequences at the termini (particularly the left terminus) and those of other beta-herpesviruses (Fig. 1b) (38–40).

The numbers of reads in the original, unfiltered data sets that matched the genome sequences were 182,459 (0.169% of the total) for EEHV1A and 74,656 (0.038%) for EEHV1B, at an average

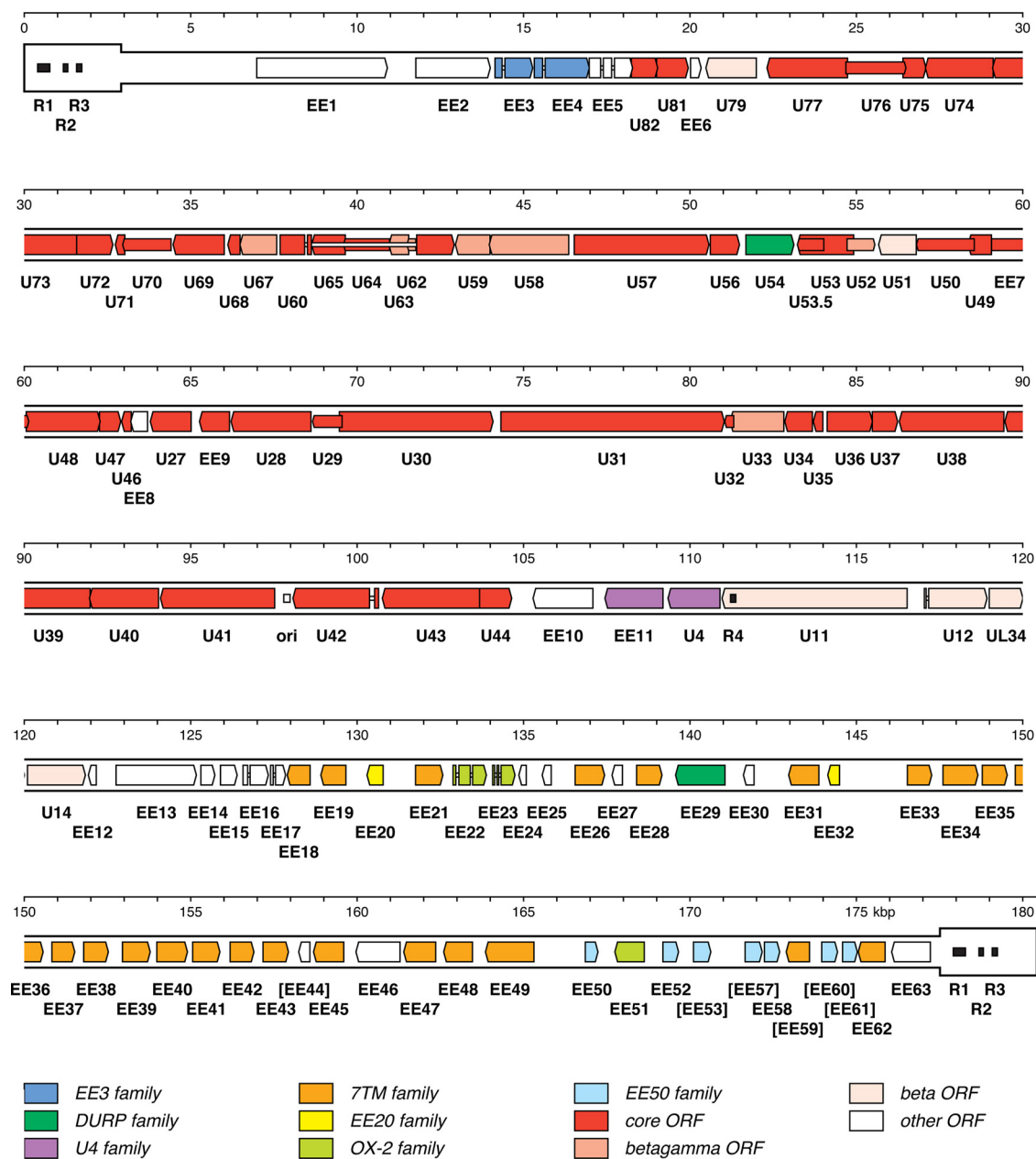


FIG 2 Map of the EEHV1A genome. TR is shown in a thicker format than U. Predicted functional ORFs are indicated by colored arrows grouped according to the key shown at the bottom, with gene nomenclature below the ORFs. Color indicates whether an ORF has an ortholog in the ancestor of the family *Herpesviridae* (core ORFs), whether it is conserved among beta- and gammaherpesviruses but not alphaherpesviruses (beta-gamma ORFs), whether it is conserved among some (but not necessarily all) betaherpesviruses (beta ORFs), whether it belongs to a paralogous family (various designations), or whether it belongs to none of these categories (other ORFs). U54 is both a beta ORF and a member of the DURP family, and U4 is both a beta ORF and a member of the U4 family. Introns are shown as narrow white bars. The names of fragmented ORFs are given in square brackets, with the ORFs depicted as intact. A putative origin of DNA replication (ori) is marked by a white-shaded rectangle, and four tandem repeats longer than the read length (R1 to R4) are marked by black-shaded rectangles.

The arrangement of ORFs near the right terminus is the same in both genomes, but the nomenclature diverges locally, for reasons explained below. Both genomes also contain four repeat regions longer than the read length (Fig. 2, R1 to R3 in TR and R4 in U) and a potential origin of DNA replication (9), which consists of two related sequences capable of forming hairpins (ori in Fig. 2). The detailed characteristics of the EEHV1A and EEHV1B ORFs are listed in Table 1 and are also available in the GenBank entries.

The locations of protein-coding regions in the EEHV1A and

EEHV1B genomes were predicted solely by bioinformatic means, as no experimental information on gene expression is available. However, ribosomal profiling data indicate that many more HCMV ORFs than the 170 recognized previously (32) are translated during infection *in vitro* (45). This finding is of unknown functional significance, but it raises the possibility that a number of small, functional, protein-coding ORFs have been missed in the analysis of the EEHV1A and EEHV1B genomes. The 751 HCMV ORFs identified by ribosome profiling may be categorized into (i)

TABLE 1 Features of predicted protein-coding regions in EEHV1A and EEHV1B

EEHV1 ORF	HCMV ortholog	Identity ^a (%)	ORF family	Protein name ^b	Protein description
EE1		95		Protein EE1	
EE2		99		Protein EE2	
EE3		32	EE3	Protein EE3	Contains signal peptide
EE4		42	EE3	Protein EE4	Contains signal peptide
EE5		51		Protein EE5	Contains signal peptide
U82	UL115	70		Envelope glycoprotein L	Contains signal peptide; complexed with envelope glycoprotein H; involved in cell entry; involved in cell-to-cell spread
U81	UL114	73		Uracil-DNA glycosylase	Involved in DNA repair
EE6 ^c				Protein EE6	Contains signal peptide
U79	UL112	89		Protein UL112	Involved in gene regulation; involved in DNA replication
U77	UL105	100		Helicase-primase helicase subunit	Involved in DNA replication
U76	UL104	100		Capsid portal protein	Dodecamer located at one capsid vertex in place of penton; involved in DNA encapsidation
U75	UL103	98		Tegument protein UL7*	Involved in virion morphogenesis
U74	UL102	98		Helicase-primase subunit	Involved in DNA replication
U73		98		DNA replication origin-binding helicase	Involved in DNA replication
U72	UL100	97		Envelope glycoprotein M	Type 3 membrane protein; 8 transmembrane domains; complexed with envelope glycoprotein N; involved in virion morphogenesis; involved in membrane fusion
U71	UL99	91		Myristylated tegument protein	Envelope-associated; involved in virion morphogenesis
U70	UL98	98		DNase	Involved in DNA processing
U69	UL97	100		Tegument serine/threonine protein kinase	Involved in protein phosphorylation
U68	UL96	100		Tegument protein UL14*	Involved in virion morphogenesis
U67	UL95	100		Protein UL95	Promotes accumulation of late transcripts
U65	UL94	96		Tegument protein UL16*	Possibly involved in virion morphogenesis
U64	UL93	99		DNA packaging tegument protein UL17*	Capsid-associated; involved in DNA encapsidation; involved in capsid transport
U63	UL92	99		Protein UL92	
U62	UL91	99		Protein UL91	
U60	UL89	99		DNA packaging terminase subunit 1	Contains ATPase domain; involved in DNA encapsidation
U59	UL88	100		Tegument protein UL88	
U58	UL87	100		Protein UL87	Promotes accumulation of late transcripts
U57	UL86	99		Major capsid protein	6 copies form hexons, 5 copies form pentons; involved in capsid morphogenesis
U56	UL85	99		Capsid triplex subunit 2	Complexed 2:1 with capsid triplex subunit 1 to connect capsid hexons and pentons; involved in capsid morphogenesis
U54	UL82/UL83/UL84	98	DURP	Protein U54	
U53.5	UL80.5	100		Capsid scaffold protein	Clipped near C terminus; involved in capsid morphogenesis
U53	UL80	100		Capsid maturation protease	Serine protease (N-terminal region); minor scaffold protein (remainder of protein, clipped near C terminus); involved in capsid morphogenesis
U52	UL79	100		Protein UL79	Promotes accumulation of late transcripts
U51	UL78	98		Envelope protein UL78	Type 3 membrane protein; 7 transmembrane domains; similar to GPCRs; putative chemokine receptor; possibly involved in intracellular signaling
U50	UL77	98		DNA packaging tegument protein UL25*	Located on capsid near vertices; possibly stabilizes the capsid and retains the genome; involved in DNA encapsidation
U49	UL76	100		Nuclear protein UL24*	
EE7		82		Thymidine kinase	Involved in nucleotide metabolism
U48	UL75	66		Envelope glycoprotein H	Type 1 membrane protein; possible membrane fusogen; complexed with envelope glycoprotein L; involved in cell entry; involved in cell-to-cell spread
U47	UL74	63		Envelope glycoprotein O	Associated with envelope glycoprotein H and envelope glycoprotein L; involved in virion morphogenesis
U46	UL73	82		Envelope glycoprotein N	Type 1 membrane protein; complexed with envelope glycoprotein M; involved in virion morphogenesis; involved in membrane fusion
EE8		74		Protein EE8	Contains potential transmembrane domain
U27	UL44	99		DNA polymerase processivity subunit	dsDNA-binding protein; involved in DNA replication
EE9		100		Ribonucleotide reductase subunit 2	Involved in nucleotide metabolism

(Continued on following page)

TABLE 1 (Continued)

EEHV1 ORF	HCMV ortholog	Identity ^a (%)	ORF family	Protein name ^b	Protein description
U28	UL45	100		Ribonucleotide reductase subunit 1	Involved in nucleotide metabolism
U29	UL46	99		Capsid triplex subunit 1	Complexed 1:2 with capsid triplex subunit 2 to connect capsid hexons and pentons; involved in capsid morphogenesis
U30	UL47	98		Tegument protein UL37*	Complexed with large tegument protein; involved in virion morphogenesis
U31	UL48	96		Large tegument protein	Complexed with tegument protein UL37; ubiquitin-specific protease (N-terminal region); involved in capsid transport
U32	UL48A	89		Small capsid protein	Located externally on capsid hexons; involved in capsid morphogenesis; possibly involved in capsid transport
U33	UL49	99		Protein UL49	
U34	UL50	100		Nuclear egress membrane protein	Type 2 membrane protein; interacts with nuclear egress lamina protein; involved in nuclear egress
U35	UL51	100		DNA packaging protein UL33*	Interacts with DNA packaging terminase subunit 2; involved in DNA encapsidation
U36	UL52	100		DNA packaging protein UL32*	Involved in DNA encapsidation; possibly involved in capsid transport
U37	UL53	100		Nuclear egress lamina protein	Interacts with nuclear egress membrane protein; involved in nuclear egress
U38	UL54	97		DNA polymerase catalytic subunit	Involved in DNA replication
U39	UL55	87		Envelope glycoprotein B	Type 1 membrane protein; possible membrane fusogen; binds cell surface heparan sulfate; involved in cell entry; involved in cell-to-cell spread
U40	UL56	100		DNA packaging terminase subunit 2	Involved in DNA encapsidation
U41	UL57	100		Single-stranded DNA-binding protein	Contains zinc finger; involved in DNA replication; possibly involved in gene regulation
U42	UL69	100		Multifunctional expression regulator	RNA-binding protein; shuttles between nucleus and cytoplasm; inhibits pre-mRNA splicing; exports virus mRNA from nucleus; exerts most effects posttranscriptionally; involved in gene regulation; involved in RNA metabolism and transport
U43	UL70	99		Helicase-primase primase subunit	Involved in DNA replication
U44	UL71	100		Tegument protein UL51*	Involved in virion morphogenesis
EE10		99		Protein EE10	
EE11		100	UL27	Protein EE11	
U4	UL27	100	UL27	Protein UL27	
U11	UL32	99		Tegument protein pp150	Major tegument protein; binds to capsids
U12	UL33	100		Envelope glycoprotein UL33	Type 3 membrane protein; 7 transmembrane domains; similar to GPCRs; putative chemokine receptor; involved in intracellular signaling
UL34	UL34	100		Protein UL34	Involved in gene regulation
U14	UL35	95		Tegument protein UL35	
EE12		94		Protein EE12	Contains potential transmembrane domain
EE13		100		Protein EE13	Contains US22 domain
EE14		100		Protein EE14	
EE15		90		Protein EE15	Contains potential transmembrane domain
EE16		62		Protein EE16	Contains signal peptide
EE17		72		Protein EE17	Contains signal peptide
EE18		99	7TM	Membrane protein EE18	Type 3 membrane protein; 7 transmembrane domains
EE19		99	7TM	Membrane protein EE19	Type 3 membrane protein; 7 transmembrane domains
EE20		100	EE20	Membrane protein EE20	Type 2 membrane protein
EE21		99	7TM	Membrane protein EE21	Type 3 membrane protein; 7 transmembrane domains; similar to GPCRs
EE22		78	OX-2	Membrane protein EE22	Type 2 membrane protein; contains 2 immunoglobulin domains; similar to OX-2 (CD200)
EE23		100	OX-2	Protein EE23	Contains signal peptide; contains immunoglobulin domain; similar to OX-2 (CD200)
EE24		100		Protein EE24	Contains potential transmembrane domain
EE25		99		Protein EE25	Contains potential transmembrane domain
EE26		94	7TM	Membrane protein EE26	Type 3 membrane protein; 7 transmembrane domains; similar to GPCRs
EE27		100		Protein EE27	
EE28		100	7TM	Membrane protein EE28	Type 3 membrane protein; 7 transmembrane domains; similar to GPCRs

(Continued on following page)

TABLE 1 (Continued)

EEHV1 ORF	HCMV ortholog	Identity ^a (%)	ORF family	Protein name ^b	Protein description
EE29		100	DURP	Protein EE29	
EE30		98		Protein EE30	
EE31		96	7TM	Membrane protein EE31	Type 3 membrane protein; 7 transmembrane domains
EE32		90	EE20	Protein EE32	Contains potential transmembrane domain
EE33		100	7TM	Membrane protein EE33	Type 3 membrane protein; 7 transmembrane domains
EE34		99	7TM	Membrane protein EE34	Type 3 membrane protein; 7 transmembrane domains; similar to GPCRs
EE35		100	7TM	Membrane protein EE35	Type 3 membrane protein; 6 transmembrane domains
EE36		99	7TM	Membrane protein EE36	Type 3 membrane protein; 7 transmembrane domains
EE37		100	7TM	Membrane protein EE37	Type 3 membrane protein; 7 transmembrane domains
EE38		100	7TM	Membrane protein EE38	Type 3 membrane protein; 7 transmembrane domains
EE39		99	7TM	Membrane protein EE39	Type 3 membrane protein; 7 transmembrane domains
EE40 ^e		97 ^g	7TM	Membrane protein EE40	Type 3 membrane protein; 7 transmembrane domains
EE41		99	7TM	Membrane protein EE41	Type 3 membrane protein; 7 transmembrane domains
EE42		100	7TM	Membrane protein EE42	Type 3 membrane protein; 7 transmembrane domains
EE43		100	7TM	Membrane protein EE43	Type 3 membrane protein; 7 transmembrane domains
EE44 ^{e,f}		75 ^g		Protein EE44	Contains signal peptide
EE45		93	7TM	Membrane protein EE45	Type 3 membrane protein; 8 transmembrane domains; similar to GPCRs
EE46		99		β -1,3-Galactosyl-O-glycosyl-glycoprotein	Contains potential transmembrane domain; involved in protein glycosylation
EE47		95	7TM	β -1,6-N-acetylglucosaminyltransferase Membrane protein EE47	Type 3 membrane protein; 8 transmembrane domains; similar to GPCRs
EE48		99	7TM	Membrane protein EE48	Type 3 membrane protein; 7 transmembrane domains
EE49		96	7TM	Membrane protein EE49	Type 3 membrane protein; 8 transmembrane domains
EE50		100	EE50	Membrane protein EE50	Contains signal peptide; contains immunoglobulin domain
EE51		97	OX-2	Membrane protein EE51	Type 2 membrane protein; contains 2 immunoglobulin domains; similar to OX-2 (CD200)
EE52		88	EE50	Membrane protein EE52	Type 2 membrane protein; contains immunoglobulin domain
EE53 ^f		45 ^g	EE50	Membrane protein EE53	Type 2 membrane protein; contains immunoglobulin domain
EE54 ^d			EE50	Membrane protein EE54	Type 2 membrane protein
EE55 ^{d,e}			EE50	Membrane protein EE55	Type 2 membrane protein
EE56 ^d			EE50	Membrane protein EE56	Type 2 membrane protein
EE57 ^{e,f}		82 ^g	EE50	Membrane protein EE57	Type 2 membrane protein; contains immunoglobulin domain
EE58 ^e		51 ^g	EE50	Membrane protein EE58	Type 2 membrane protein
EE59 ^{f,h}			7TM	Membrane protein EE59	Type 3 membrane protein; 7 transmembrane domains
EE60 ^{e,f}			EE50	Membrane protein EE60	Type 2 membrane protein; contains immunoglobulin domain
EE61 ^{e,f}			EE50	Membrane protein EE61	Type 2 membrane protein
EE62 ^e			7TM	Membrane protein EE62	Type 3 membrane protein; 7 transmembrane domains
EE63		58		α -(1,3)-Fucosyltransferase	Contains potential transmembrane domain; involved in protein glycosylation

^a Calculated from EEHV1A and EEHV1B amino acid sequences aligned by using GCG Gap and shown to the nearest integer.

^b The standard nomenclature for orthologous herpesvirus proteins employed in NCBI Reference Sequence files is used for conserved ORFs. In this system, the names of proteins specified by certain ORFs are derived from HCMV nomenclature (e.g., tegument protein UL35, encoded by U14). However, the names of proteins specified by certain core ORFs are derived from nomenclature of the alphaherpesvirus herpes simplex virus type 1 (e.g., DNA packaging tegument protein UL25, encoded by U50). Names in the latter category may cause confusion in relation to HCMV ORF nomenclature and are marked by asterisks.

^c ORF absent from EEHV1B.

^d ORF absent from EEHV1A.

^e Fragmented in EEHV1B.

^f Fragmented in EEHV1A.

^g Repaired, nonfragmented sequences used.

^h EEHV1A ORF too fragmented to be repaired.

148 previously recognized (canonical) ORFs, (ii) 77 ORFs corresponding to forms of the canonical ORFs that are extended or truncated at their 5' ends, (iii) 14 ORFs containing parts of canonical ORFs due to alternative splicing, and (iv) 512 ORFs not containing any parts of canonical ORFs. A large number of these ORFs employ nonconventional initiation codons, and many are small, with some encoding a single amino acid residue. In order to determine which are conserved in EEHV1, TBLASTN searches of

the EEHV1A and EEHV1B genomes were conducted using parameters tailored to ORF length. The expectation threshold applied was <0.05 , a level at which most ORFs in categories i and ii that are considered to have EEHV1 orthologs scored positive. Of the 526 ORFs in categories iii and iv, only 14 scored positive, and the significance of only two of these was supported by positional conservation. The first, ORF245C (86 codons), is antiparallel to the 5' end of canonical UL105 (orthologous to U77 in EEHV1)

and exhibits similarity for 69 codons, and the second, ORF220C (51 codons), is antiparallel to the 3' end of canonical UL92 (orthologous to U63 in EEHV1) and exhibits similarity for 26 codons. However, the amino acid sequences of the corresponding regions of U77 and U63 are more highly conserved than those of ORF245C and ORF220C, and these regions of U77 and U63 are also highly conserved among all herpesviruses that possess orthologs. Moreover, ORF245C is fragmented by stop codons in HHV6A, HHV6B, and HHV7. Thus, it is likely that conservation of ORF245C and ORF220C is due indirectly to conservation of U77 and U63 encoded on the antiparallel strand. In summary, we obtained no convincing evidence for direct functional conservation in EEHV1 of any of the HCMV ORFs in categories iii and iv. Also, in instances where a canonical ORF in category i could be distinguished in the analytical output from its 5'-extended or 5'-truncated counterparts in category ii, no changes to the predicted 5' ends of EEHV1 ORFs were indicated.

Nonetheless, the EEHV1 map may need to be refined in light of future experimental data, as it is conceivable that some protein-coding ORFs that are not conserved in HCMV have been missed or that some of the ORFs predicted to encode functional proteins may turn out not to do so. Splice sites were predicted sparingly and tentatively (in 10 genes), under conditions in which their use in both genomes would significantly extend coding capacity. In reality, splicing may be more common. Genes specifying non-translated RNAs, including miRNAs processed from longer RNAs, would not have been detected.

Nomenclature of protein-coding regions. For EEHV1 ORFs that have detectable orthologs in other betaherpesviruses, we retained the established nomenclature (9), which is based on that of HHV6A (U82, etc.) and is implicit in Fig. 2 and Table 1. This does not imply that EEHV1 is phylogenetically more closely related to HHV6A than any other betaherpesvirus (see below). The names of orthologs in HCMV are also listed in Table 1. We named one ORF (UL34) after its ortholog in HCMV, because a counterpart is lacking in HHV6A. One category of coding regions in the orthologous class (termed core ORFs in Fig. 2) is perceived as having been present in the ancestor of all members of the family *Herpesviridae* (i.e., alpha-, beta-, and gammaherpesviruses) (46). A second category (termed beta-gamma ORFs) has orthologs in beta- and gamma- but not alphaherpesviruses, and a third category (termed beta ORFs) has counterparts in some or all betaherpesviruses but not alpha- or gammaherpesviruses. Orthology was assigned largely on the basis of the highest BLASTP score being achieved by a betaherpesvirus ORF when all NCBI reference proteins were searched using an EEHV1A or EEHV1B ORF. In addition, orthology was assigned to several ORFs (U71, U68, U62, U53.5, U28, U11, and U12) that did not meet the above criterion, on the basis of the highest BLASTP score being achieved by a positionally equivalent betaherpesvirus ORF when all NCBI herpesvirus reference proteins were searched. Orthology was also assigned to a few ORFs (U79, U75, U54, U51, U47, and U32) that did not meet either of the above criteria, on the basis of the amino acid sequence of an EEHV1A or EEHV1B ORF sharing a conserved motif or other distinguishable feature with a positionally equivalent betaherpesvirus ORF. Among these ORFs, two names were assigned particularly tentatively. That of U47 (encoding glycoprotein O) was adopted on the basis of the presence of diagnostic cysteine residues, as argued previously (9). The name U54 (encoding a member of the DURP family) could not be assigned unambigu-

ously, because other betaherpesviruses have several paralogous ORFs at this location (e.g., U54 and U55 in HHV6A and UL82, UL83, and UL84 in HCMV). Thus, U54 in EEHV1A and EEHV1B could equally well have been named U55. As a result, the function of this ORF could not be predicted.

ORFs that lack detectable orthologs in other betaherpesviruses were named EE1 to EE63 and number 60 in EEHV1A and 59 in EEHV1B. EEHV1A lacks EE54, EE55, and EE56, and EEHV1B lacks EE6, EE60, EE61, and EE62. It is striking that the majority of ORFs lacking orthologs are predicted to encode membrane-associated proteins. In some instances, naming of these ORFs took on complex aspects. EE11 is related to U4, and the orthologous name was applied to the member of this paralogous pair that is most closely related to HHV6A U4. However, as in the case of U54 discussed above, these ORFs could have been named the other way around. U4 paralogs also feature in HHV6A (41) and some other betaherpesviruses (47).

The ORFs that lack detectable orthologs in other betaherpesviruses include some that have relatives in alpha- and gammaherpesviruses. EE7 and EE9, which encode thymidine kinase and ribonucleotide reductase subunit 2, respectively (9), are core ORFs that are conserved positionally in alpha- and gammaherpesviruses but have been lost in other betaherpesvirus lineages. EE46 encodes an enzyme involved in protein glycosylation, namely, β -1,3-galactosyl-O-glycosyl-glycoprotein β -1,6-N-acetylglucosaminyltransferase, as does the gammaherpesvirus bovine herpesvirus 4 (48). The EEHV1 and bovine herpesvirus 4 ORFs are likely to have originated by independent gene capture, since the viruses concerned belong to different subfamilies, and the latter version is very closely related to its bovine counterpart (49). EE63 encodes a second enzyme involved in protein glycosylation, namely, α -(1,3)-fucosyltransferase, but lacks orthologs in other herpesviruses. The proteins encoded by EE46 and EE63 are both related closely to their host counterparts (as assessed via data for the African elephant). The region containing EE1 to EE63 also includes a few ORFs that have relatives in other betaherpesviruses but are likely to have resulted from independent gene capture events. Some encode proteins that are related to cellular OX-2 (OX-2 family, with EE51 being very closely related to its African elephant counterpart), G protein-coupled receptors (GPCRs; some members of the 7TM family), or immunoglobulin domain-containing proteins (some members of the EE50 family).

Arrangement of protein-coding regions. The layout of ORFs in the EEHV1A and EEHV1B genomes may be considered in terms of three separate sections. The section extending from U82 to U14 contains 56 ORFs that have orthologs in other betaherpesviruses, two ORFs (EE7 and EE9) that have orthologs in alpha- and gammaherpesviruses, and three ORFs (EE8, EE10, and EE11) that lack counterparts in other herpesviruses. Notably, the arrangement of ORFs that have betaherpesvirus orthologs is not colinear with that in other betaherpesviruses. Instead, these ORFs are present in three rearranged blocks (U82-U46, U27-U44, and U4-U14). Taking this layout into account, EE8 is positionally similar to ORFs in other betaherpesviruses that, like EE8, encode proteins containing potential transmembrane domains (U20 to U24 in HHV6A and UL40, UL41A, and UL42 in HCMV). However, this indication of possible orthology is too slight, and divergence between ORFs at this location in other betaherpesviruses is too great to make such an assignment with any confidence. Also, EE10 is positionally equivalent to a core ORF in other betaherpesviruses

that is a member of the DURP family (U45 in HHV6A and UL72 in HCMV, considered to have been derived from a captured deoxyuridine triphosphatase gene and to have lost regions responsible for catalytic activity [46]). However, the encoded protein lacks a detectable relationship with DURP family members, making such an assignment unjustified. It is notable that orthologs of a group of three HCMV genes (UL128, UL130, and UL131A) that are involved prominently in endothelial cell tropism (50) are absent from EEHV1.

The section of the genome near the left terminus contains EE1 to EE5. Three of these ORFs (EE3 and EE4, which belong to the EE3 family, and EE5) are similar in encoding proteins that contain signal peptides and are rich in S or T residues. Among potential positional counterparts in other betaherpesviruses, UL116 in HCMV and other members of the genus *Cytomegalovirus* has similar properties, but amino acid sequence similarity is negligible. The leftmost ORFs in the EEHV1A and EEHV1B genomes, EE1 and EE2, are equivalent in location and size to the major immediate-early transcriptional activator ORFs of other betaherpesviruses (U90 and U86 in HHV6A and UL123 and UL122 in HCMV, respectively), but again there is not enough sequence similarity to make assignments of orthology. However, the region that encompasses a substantial sequence upstream from EE1, the whole of EE1, the sequence between EE1 and EE2, and the 5' end of EE2 (approximately nt 5500 to 12000 in EEHV1A) is significantly depleted in the CG dinucleotide, at 43% of that in the rest of the genome, despite having a higher G+C content (49%). CG depletion is a hallmark of the major immediate-early genes in other betaherpesviruses (51).

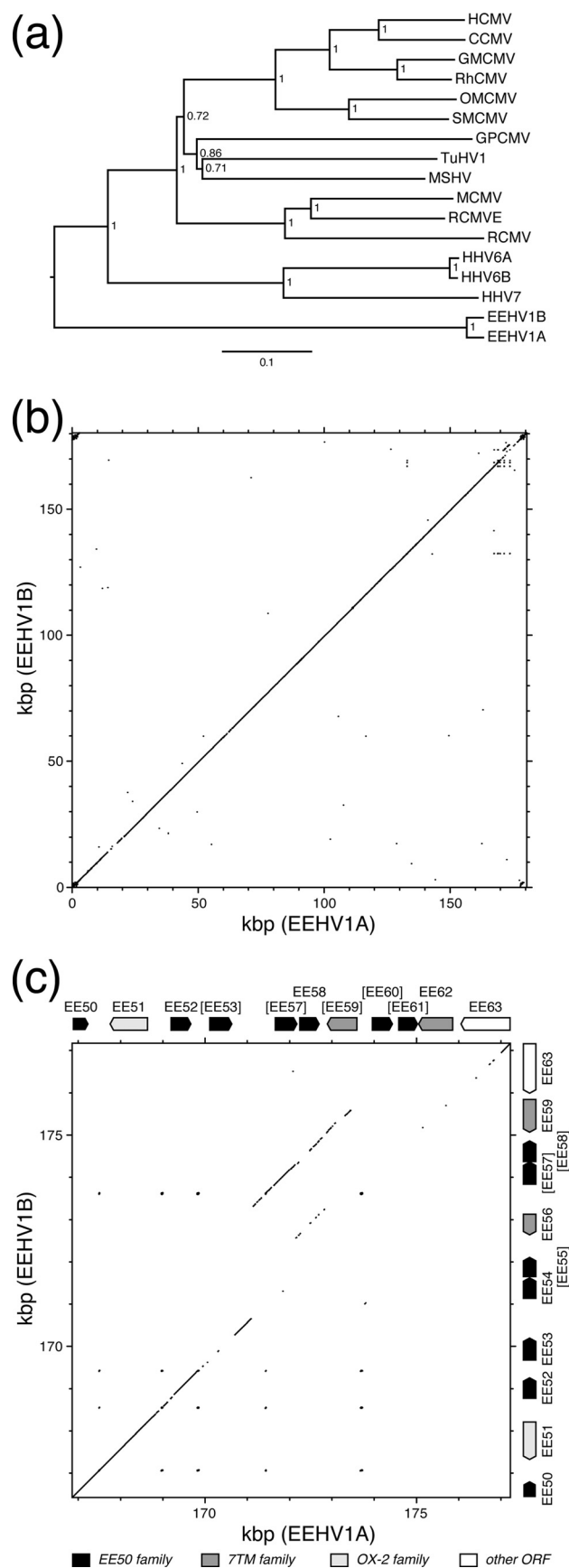
The remaining section of the genome occupies the one-third near the right terminus and consists of EE12 to EE63. Among these ORFs, which number 49 in each genome, are 23 assigned to the 7TM family, because the encoded proteins contain seven predicted transmembrane domains (although one contains six such domains and three contain eight; Table 1). The 7TM family was defined on a wider basis than the other families, in that it includes all ORFs that lack orthologs in other betaherpesviruses and encode 7TM proteins (a class of type 3 membrane protein), regardless of whether amino acid sequence similarity was detected. Even in instances where similarity was detected, such as among the EE33 to EE43 set of tandem ORFs, it is low. Some members of the 7TM family are related to cellular GPCRs (Table 1), indicating that they originated from a captured host GPCR gene. However, given the range of relationships that were detected, it is possible that the 7TM family arose from more than one source rather than via successive duplications of a single gene. It is notable that, among the other betaherpesviruses, HCMV has two families of ORFs specifying 7TM proteins. These are the US12 family, which consists of US12-US21, and the GPCR family, which consists of US27, US28, UL33, and UL78 (41, 52). Some members of the US12 family may have roles in virion maturation and egress (53), and some members of the GPCR family, or their orthologs in other betaherpesviruses, function in intracellular signaling (54–56). It has been suggested that the HCMV US12 and GPCR families are related to each other, but there is little evidence for this from sequence alignments (57, 58). In addition to the 7TM family, the EEHV1A and EEHV1B genomes contain three other ORFs encoding 7TM proteins, namely, two members of the GPCR family (U51 and U12, orthologs of HCMV UL78 and UL33, respectively) and U72 (which has eight transmembrane domains). How-

ever, these ORFs have orthologs in other herpesviruses and are likely to have more ancient origins in the EEHV lineage than does the 7TM family.

Also in the section containing EE12 to EE63 are two ORFs belonging to the EE20 family and three ORFs belonging to the OX-2 family. There is no convincing evidence that any of the latter is orthologous to the OX-2-like ORFs present in other betaherpesviruses (e.g., U85 in HHV-6A [41] and UL119 in HCMV [59]). Seven ORFs near the right genome terminus belong to the EE50 family and are predicted to encode (or to have once encoded) type 2 membrane proteins, some of which contain immunoglobulin domains.

Relationships between the genomes and integration of previously published sequence data. A considerable amount of sequence information was already available for EEHV1A and EEHV1B strains during our analysis. The EEHV1A genome sequence was confirmed as being most closely related to the former and that of EEHV1B to the latter. The largest sequence yet published (AF322977, from an EEHV1B strain) is 59,467 bp in size (9) and corresponds to two separate regions of the EEHV1A and EEHV1B genomes, consisting of U70-U46 (equivalent to nt 32929 to 63190 in EEHV1A) joined to U11-U38 in inverse orientation (nt 86078 to 115551). This atypical arrangement is probably the result of an assembly error in the previously published sequence, which was generated via demanding PCR experiments. The sequences of several substantial regions from EEHV1 strains have been deposited in GenBank by L. K. Richman and colleagues, the largest originating from EEHV1A and containing U77-U58 (HM568515; nt 23873 to 46119 in EEHV1A) from one strain and U58-U29 (HM568525; nt 45799 to 69324 in EEHV1A) from another strain. The latter sequence fully supports our conclusions about a putative assembly error in AF322977. Richman and colleagues have also contributed the sequences of shorter regions from other genome locations, including parts of U31 (e.g., JN983083) and U33 (e.g., JN983082) and the whole of EE1 (JX011081). In addition, they and several other groups have deposited in GenBank many sequences from various EEHV strains that are contained within the longer regions described above.

A phylogenetic tree was constructed by using MEGA v. 5.03 (60) on the basis of a concatenation of the amino acid sequences of several highly conserved betaherpesvirus ORFs and shows that EEHV1A and EEHV1B are related closely to each other and form a clade outlying the other betaherpesviruses (Fig. 3a). This finding is in accord with previous reports based on shorter sequences and demonstrates the origins of the genus *Proboscivirus* from the earliest known branching event in subfamily *Betaherpesvirinae* (6, 8). The two other main branches represent the genus *Roseolovirus* and the genera *Cytomegalovirus* and *Muromegalovirus* together with the unclassified betaherpesviruses. Although EEHV1 and the members of genus *Roseolovirus* share a similar genome structure, retain a core gene (U73) that has been lost in the other betaherpesvirus genera, and feature a similar scheme for ORF nomenclature, EEHV1 is no more closely related phylogenetically to the members of genus *Roseolovirus* than it is to members of the other genera (Fig. 3a). The events that led to the differences in ORF order in EEHV1 probably occurred after branching of genus *Proboscivirus* from the other betaherpesviruses and before the latter diverged further into the modern genera. The ORF arrangement in the ancestral virus cannot be discerned, but it seems likely that inversion of a segment containing U27-U44, which resulted in the



wide separation of U44 from U46, occurred in the *Proboscivirus* lineage, since the orthologs of these two ORFs are located close to each other in all other members of the family *Herpesviridae*.

A DNA sequence comparison shows that the EEHV1A and EEHV1B genomes are highly colinear throughout their lengths (Fig. 3b). This finding was not an artifact of having assumed colinearity during derivation of the EEHV1B sequence, as the integrity of both sequences had been confirmed by assembling the original, unfiltered data sets against the finished sequences. From inspection of Fig. 3b, the regions that are most obviously divergent are located near the left terminus approximately at nt 14,000 to 20,000 and near the right terminus at nt 170,000 to 177,000.

Further resolution of the divergence of the EEHV1A and EEHV1B genomes is provided by the values for amino acid sequence identity listed in Table 1 for the pairs of ORFs. In the region near the left terminus, EE3 and EE4 are the first and second most divergent ORFs, respectively, in the genome, and the adjacent EE5 is the fourth most divergent. In the region near the right terminus, two members of the EE50 family (EE53 and EE58) are the third and fifth most divergent ORFs, respectively. At the other end of the spectrum, the amino acid sequences of over half of the EEHV1A and EEHV1B ORFs (64 pairs) are $\geq 99\%$ identical to each other. Many ORFs that have orthologs in other herpesviruses are in this highly conserved group, but several not in this category are also included. The pattern of divergence between the EEHV1A and EEHV1B genomes is somewhat reminiscent of that between HHV6A and HHV6B (43). Thus, the regions of most extensive difference between HHV6A and HHV6B tend to be located near the genome termini.

In addition to exhibiting extensive sequence divergence, a higher resolution comparison of the DNA sequences near the right genome terminus demonstrates a degree of noncolinearity (Fig. 3c). Similarity in the region containing EE57 to EE59 is displaced from the diagonal, indicating that EEHV1B contains an insertion to the right of EE53 relative to EEHV1A and that EEHV1A contains an insertion to the left of EE63 relative to EEHV1B. Thus, although on first glance the ORF arrangement in this region seems to be the same in both genomes, the orthologous relationships are more complex and may represent the occurrence of duplications and deletions. This prompted a variation in nomenclature in this region indicating that EE54 to EE56 (of which one is fragmented) are present in EEHV1B but not EEHV1A and that EE60 to EE62 (of which two are fragmented) are present in EEHV1A but not EEHV1B (Fig. 3c).

FIG 3 Comparisons of the EEHV1A and EEHV1B genomes. (a) Phylogenetic analysis of the concatenated amino acid sequences of U38, U39, U40, U41, U57, U60, U77, and U81 from EEHV1A and EEHV1B and their orthologs in other betaherpesviruses. Abbreviations are given in the legend to Fig. 1 and also include the following: RhCMV, rhesus cytomegalovirus; TuHV1, tupaia herpesvirus 1; MSHV, *Miniopterus schreibersii* herpesvirus; and RCMVE, rat cytomegalovirus England. The tree was constructed by using the neighbor-joining method, rooting at the midpoint. Confidence levels were calculated by using bootstrapping (2,000 replicates) and are shown as fractions. The scale shows nucleotide differences/nucleotide. (b) Matrix sequence comparison plot of the complete EEHV1A and EEHV1B genomes. (c) Matrix sequence comparison plot of the regions near the right terminus of the EEHV1A and EEHV1B genomes, corresponding to an expansion of part of the upper portion of the plot shown in panel b. The layout of ORFs in each sequence is illustrated, with shading indicating ORF families provided in the key at the bottom. The plots in panels b and c were computed by using GCG Compare and GCG Dotplot (window, 25; stringency, 21).

Fragmented protein-coding regions. Six ORFs in EEHV1A (EE44, EE53, EE57, EE59, EE60, and EE61) and five ORFs in EEHV1B (EE40, EE44, EE55, EE57, and EE58), all of which are located toward the right terminus, were assessed as being fragmented. For the associated analyses, the sequences of these ORFs were repaired conceptually by making minimal adjustments, which, of course, are not included in the GenBank entries. Of the fragmented ORFs, all but EE40, EE44, and EE61 are members of the EE50 family. It is notable that several fragmented ORFs (EE44, EE53, EE60, and EE61 in EEHV1A and EE44, EE55, and EE58 in EEHV1B) are frameshifted in a G:C mononucleotide tract. Other ORFs are frameshifted in other sequences (EE40 and EE57 in EEHV1B), contain an in-frame stop codon (EE57 in EEHV1A), or have multiple mutations (EE59 in EEHV1A and EE57 in EEHV1B). Additional ORFs in this region (EE58, EE62, and EE63 in EEHV1A and EE53, EE54, EE56, EE59, and EE63 in EEHV1B) also contain G:C tracts but are not frameshifted. However, the sequences at G:C tracts were analyzed in terms of mode lengths, and inherent variability at any of these locations could result in some genomes in which the cognate ORF is frameshifted and others in which it is not. This situation is reminiscent of the G:C tract-mediated frameshifting evident in the highly variable RL11 family of HCMV and other primate cytomegaloviruses, which has been suggested as a means of regulating certain functions via selection of viral populations (61). The observed relationships between EEHV1A and EEHV1B, and the presence of fragmented ORFs in this region of the genome, suggest that members of the EE50 family are under intense evolutionary selection, with constituent ORFs having appeared and disappeared, as well as having diverged rapidly while present. Similar complex phenomena have been described for ORF families in HCMV and other primate cytomegaloviruses (53, 62).

Concluding remarks. By using high-throughput methods, we have determined the complete genome sequences of the two most lethal EEHV genotypes directly from fatal cases. The data sets contained tiny proportions of viral reads among an abundance of cellular reads from a host whose genome sequence was not known. The availability of the viral genome sequences will make assembling the sequences of additional strains of EEHV1A and EEHV1B, and perhaps other EEHV genotypes, much easier. This, in turn, will aid work on the diagnosis and epidemiology of EEHV1-associated hemorrhagic disease. The sequences will also inform research on EEHV1 genes, and perhaps resource novel ways of attempting to grow virus in cell culture. However, the sequences give no obvious clues to which viral genes may be key to pathogenesis, particularly the marked endothelial cell tropism that is evident during infection. Nonetheless, genes that are not conserved in other betaherpesviruses and for which functions have been predicted, the impressively large number of members of the 7TM family, and the evolutionary variability of the EE50 family may provide areas of immediate interest.

ACKNOWLEDGMENTS

This work was supported by the Zoological Society London, the Department for Environment Food and Rural Affairs, the Animal Health and Veterinary Laboratories Agency, the UK Biotechnology and Biological Sciences Research Council (BB/J004243/1, BB/J004235/1, and BB/J004324/1), and the UK Medical Research Council.

We thank Derek Gatherer (MRC–University of Glasgow Centre for Virus Research) for extensive advice on bioinformatics and Wai Kwong

Lee and Andrew Carswell (BHF Glasgow Cardiovascular Research Centre, University of Glasgow) for providing Sanger DNA sequencing services.

REFERENCES

1. Zoos Forum Elephant Working Group. 2010. Elephants in UK zoos. Defra, Bristol, United Kingdom. <http://archive.defra.gov.uk/wildlife-pets/zoos/documents/elephant-forum-1007.pdf>.
2. Anonymous. 2011. With few resources, researchers work to contain fatal elephant virus. *Am. J. Vet. Res.* 72:1006.
3. Hayward GS. 2012. Conservation: clarifying the risk from herpesvirus to captive Asian elephants. *Vet. Rec.* 170:202–203.
4. Ossent P, Guscetti F, Metzler AE, Lang EM, Rübel A, Hauser B. 1990. Acute and fatal herpesvirus infection in a young Asian elephant (*Elephas maximus*). *Vet. Pathol.* 27:131–133.
5. Richman LK, Montali RJ, Cambre RC, Schmitt D, Hardy D, Hildebrandt T, Bengis RG, Hamzeh FM, Shahkolahi A, Hayward GS. 2000. Clinical and pathological findings of a newly recognized disease of elephants caused by endotheliotropic herpesviruses. *J. Wildl. Dis.* 36:1–12.
6. Richman LK, Montali RJ, Garber RL, Kennedy MA, Lehnhardt J, Hildebrandt T, Schmitt D, Hardy D, Alcendor DJ, Hayward GS. 1999. Novel endotheliotropic herpesviruses fatal for Asian and African elephants. *Science* 283:1171–1176.
7. Fickel J, Richman LK, Montali R, Schaftenaar W, Göritz F, Hildebrandt TB, Pitra C. 2001. A variant of the endotheliotropic herpesvirus in Asian elephants (*Elephas maximus*) in European zoos. *Vet. Microbiol.* 82:103–109.
8. Ehlers B, Burkhardt S, Goltz M, Bergmann V, Ochs A, Weiler H, Hentschke J. 2001. Genetic and ultrastructural characterization of a European isolate of the fatal endotheliotropic elephant herpesvirus. *J. Gen. Virol.* 82:475–482.
9. Ehlers B, Dural G, Marschall M, Schregel V, Goltz M, Hentschke J. 2006. Endotheliotropic elephant herpesvirus, the first betaherpesvirus with a thymidine kinase gene. *J. Gen. Virol.* 87:2781–2789.
10. Pellett PE, Davison AJ, Eberle R, Ehlers B, Hayward GS, Lacoste V, Minson CA, Nicholas J, Roizman B, Studdert MJ, Wang F. 2011. *Herpesvirales*, p 99–107. In King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (ed), *Virus taxonomy*, ninth report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, London, United Kingdom.
11. Garner MM, Helmick K, Ochsenreiter J, Richman LK, Latimer E, Wise AG, Maes RK, Kiupel M, Nordhausen RW, Zong JC, Hayward GS. 2009. Clinico-pathologic features of fatal disease attributed to new variants of endotheliotropic herpesviruses in two Asian elephants (*Elephas maximus*). *Vet. Pathol.* 46:97–104.
12. Latimer E, Zong J-C, Heaggans SY, Richman LK, Hayward GS. 2011. Detection and evaluation of novel herpesviruses in routine and pathological samples from Asian and African elephants: identification of two new probosciviruses (EEHV5 and EEHV6) and two new gammaherpesviruses (EGHV3B and EGHV5). *Vet. Microbiol.* 147:28–41.
13. Denk D, Stidworthy MF, Redrobe S, Latimer E, Hayward GS, Cracknell J, Claessens A, Steinbach F, McGowan S, Dastjerdi A. 2012. Fatal elephant endotheliotropic herpesvirus type 5 infection in a captive Asian elephant. *Vet. Rec.* 171:380–381.
14. McCully RM, Basson PA, Pienaar JG, Erasmus BJ, Young E. 1971. Herpes nodules in the lung of the African elephant (*Loxodonta africana* (Blumebach, 1792)). *Onderstepoort J. Vet. Res.* 38:225–235.
15. Jacobson ER, Sundberg JP, Gaskin JM, Kollias GV, O'Banion MK. 1986. Cutaneous papillomas associated with a herpesvirus-like infection in a herd of captive African elephants. *J. Am. Vet. Med. Assoc.* 189:1075–1078.
16. Ryan SJ, Thompson SD. 2001. Disease risk and inter-institutional transfer of specimens in cooperative breeding programs: herpes and the elephant species survival plans. *Zoo Biol.* 20:89–101.
17. Reid CE, Hildebrandt TB, Marx N, Hunt M, Thy N, Reynes JM, Schaftenaar W, Fickel J. 2006. Endotheliotropic elephant herpes virus (EEHV) infection. The first PCR-confirmed fatal case in Asia. *Vet. Q.* 28:61–64.
18. Stanton JJ, Zong J-C, Latimer E, Tan J, Herron A, Hayward GS, Ling PD. 2010. Detection of pathogenic elephant endotheliotropic herpesvirus in routine trunkwashes from healthy adult Asian elephants (*Elephas maximus*) by use of a real-time quantitative polymerase chain reaction assay. *Am. J. Vet. Res.* 71:925–933.
19. Schaftenaar W, Reid C, Martina B, Fickel J, Osterhaus AD. 2010.

- Nonfatal clinical presentation of elephant endotheliotropic herpes virus discovered in a group of captive Asian elephants (*Elephas maximus*). *J. Zoo. Wildl. Med.* 41:626–632.
20. Hardman K, Dastjerdi A, Gurrall R, Routh A, Banks M, Steinbach F, Bouts T. 2012. Detection of elephant endotheliotropic herpesvirus type 1 in asymptomatic elephants using TaqMan real-time PCR. *Vet. Rec.* 170: 205.
 21. Schmitt DL, Hardy DA, Montali RJ, Richman LK, Lindsay WA, Isaza R, West G. 2000. Use of famciclovir for the treatment of endotheliotropic herpesvirus infections in Asian elephants (*Elephas maximus*). *J. Zoo. Wildl. Med.* 31:518–522.
 22. Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589–595.
 23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
 24. Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J. 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.* 18:802–809.
 25. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18:821–829.
 26. Warren RL, Sutton GG, Jones SJM, Holt RA. 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 23:500–501.
 27. Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using Phred I. Accuracy assessment. *Genome Res.* 8:175–185.
 28. Ewing B, Green P. 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* 8:186–194.
 29. Tsai IJ, Otto TD, Berriman M. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* 11:R41.
 30. Staden R, Beal KF, Bonfield JK. 2000. The Staden package, 1998. *Methods Mol. Biol.* 132:115–130.
 31. Cunningham C, Gatherer D, Hilfrich B, Baluchova K, Dargan DJ, Thomson M, Griffiths PD, Wilkinson GWG, Schulz TF, Davison AJ. 2010. Sequences of complete human cytomegalovirus genomes from infected cell cultures and clinical specimens. *J. Gen. Virol.* 91:605–615.
 32. Gatherer D, Seirafian S, Cunningham C, Holton M, Dargan DJ, Baluchova K, Hector RD, Galbraith J, Herzyk P, Wilkinson GWG, Davison AJ. 2011. High-resolution human cytomegalovirus transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* 108:19755–19760.
 33. Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18:1851–1858.
 34. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D. 2010. Tablet—next generation sequence assembly visualization. *Bioinformatics* 26:401–402.
 35. Davison AJ, Dolan A, Akter P, Addison C, Dargan DJ, Alcendor DJ, McGeoch DJ, Hayward GS. 2003. The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. *J. Gen. Virol.* 84:17–28.
 36. Gao M, Isom HC. 1984. Characterization of the guinea pig cytomegalovirus genome by molecular cloning and physical mapping. *J. Virol.* 52: 436–447.
 37. Tamashiro JC, Spector DH. 1986. Terminal structure and heterogeneity in human cytomegalovirus strain AD169. *J. Virol.* 59:591–604.
 38. Deiss LP, Chou J, Frenkel N. 1986. Functional domains within the *a* sequence involved in the cleavage-packaging of herpes simplex virus DNA. *J. Virol.* 59:605–618.
 39. Hammerschmidt W, Ludwig H, Buhk H-J. 1988. Specificity of cleavage in replicative-form DNA of bovine herpesvirus 1. *J. Virol.* 62:1355–1363.
 40. Thomson BJ, Dewhurst S, Gray D. 1994. Structure and heterogeneity of the *a* sequences of human herpesvirus 6 strain variants U1102 and Z29 and identification of human telomeric repeat sequences at the genomic termini. *J. Virol.* 68:3007–3014.
 41. Gompels UA, Nicholas J, Lawrence G, Jones M, Thomson BJ, Martin MED, Efsthathiou S, Craxton M, Macaulay HA. 1995. The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution. *Virology* 209:29–51.
 42. Nicholas J. 1996. Determination and analysis of the complete nucleotide sequence of human herpesvirus. *J. Virol.* 70:5975–5989.
 43. Dominguez G, Dambaugh TR, Stamey FR, Dewhurst S, Inoue N, Pellett PE. 1999. Human herpesvirus 6B genome sequence: coding content and comparison with human herpesvirus 6A. *J. Virol.* 73:8040–8052.
 44. Rixon FJ. 1993. Structure and assembly of herpesviruses. *Semin. Virol.* 4:135–144.
 45. Stern-Ginossar N, Weisburd B, Michalski A, Le VTK, Hein MY, Huang S-X, Ma M, Shen B, Qian S-B, Hengel H, Mann M, Ingolia NT, Weissman JS. 2012. Decoding human cytomegalovirus. *Science* 338: 1088–1093.
 46. McGeoch DJ, Rixon FJ, Davison AJ. 2006. Topics in herpesvirus genomics and evolution. *Virus Res.* 117:90–104.
 47. Zhang H, Todd S, Tachedjian M, Barr JA, Luo M, Yu M, Marsh GA, Cramer G, Wang L-F. 2012. A novel bat herpesvirus encodes homologues of major histocompatibility complex classes I and II, C-type lectin, and a unique family of immune-related genes. *J. Virol.* 86:8014–8030.
 48. Zimmermann W, Broll H, Ehlers B, Buhk H-J, Rosenthal A, Goltz M. 2001. Genome sequence of bovine herpesvirus 4, a bovine *Rhadinovirus*, and identification of an origin of DNA replication. *J. Virol.* 75:1186–1194.
 49. Markine-Goriaynoff N, Georgin J-P, Goltz M, Zimmermann W, Broll H, Wamwayi HM, Pastoret P-P, Sharp PM, Vanderplassen A. 2003. The core 2 β -1,6-*N*-acetylglucosaminyltransferase-mucin encoded by bovine herpesvirus 4 was acquired from an ancestor of the African buffalo. *J. Virol.* 77:1784–1792.
 50. Wang D, Shenk T. 2005. Human cytomegalovirus virion protein complex required for epithelial and endothelial cell tropism. *Proc. Natl. Acad. Sci. U. S. A.* 102:18153–18158.
 51. Honess RW, Gompels UA, Barrell BG, Craxton M, Cameron KR, Staden R, Chang Y-N, Hayward GS. 1989. Deviations from expected frequencies of CpG dinucleotides in herpesvirus DNAs may be diagnostic of differences in the states of their latent genomes. *J. Gen. Virol.* 70:837–855.
 52. Chee MS, Bankier AT, Beck S, Bohni R, Brown CM, Cerny R, Horsnell T, Hutchison CA, III, Kourazides T, Martignetti JA, Preddie E, Satchwell SC, Tomlinson P, Weston KM, Barrell BG. 1990. Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. *Curr. Top. Microbiol. Immunol.* 154:125–169.
 53. Das S, Pellett PE. 2007. Members of the HCMV US12 family of predicted heptaspanning membrane proteins have unique intracellular distributions, including association with the cytoplasmic virion assembly complex. *Virology* 361:263–273.
 54. Neote K, DiGregorio D, Mak JY, Horuk R, Schall TJ. 1993. Molecular cloning, functional expression, and signaling characteristics of a C-C chemokine receptor. *Cell* 72:415–425.
 55. Isegawa Y, Ping Z, Nakano K, Sugimoto N, Yamanishi K. 1998. Human herpesvirus 6 open reading frame U12 encodes a functional β -chemokine receptor. *J. Virol.* 72:6104–6112.
 56. Milne RSB, Mattick C, Nicholson L, Devaraj P, Alcamí A, Gompels UA. 2000. RANTES binding and down-regulation by a novel human herpesvirus-6 β chemokine receptor. *J. Immunol.* 164:2396–2404.
 57. Rigoutsos I, Novotny J, Huynh T, Chin-Bow ST, Parida L, Platt D, Coleman D, Shenk T. 2003. In silico pattern-based analysis of the human cytomegalovirus genome. *J. Virol.* 77:4326–4344.
 58. Lesniewski M, Das S, Skomorowska-Prokvolit Y, Wang F-Z, Pellett PE. 2006. Primate cytomegalovirus US12 gene family: a distinct and diverse clade of seven-transmembrane proteins. *Virology* 354:286–298.
 59. Davison AJ, Bhella D. 2007. Comparative genome and virion structure, p 177–203. In Arvin A, Campadelli-Fiume G, Mocarski E, Moore PS, Roizman B, Whitley R, Yamanishi K (ed), *Human herpesviruses: biology, therapy and immunophylaxis*. Cambridge University Press, Cambridge, United Kingdom.
 60. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28:2731–2739.
 61. Davison AJ, Holton M, Dolan A, Dargan DJ, Gatherer D, Hayward GS. Comparative genomics of primate cytomegaloviruses. In Reddehase MJ (ed), *Cytomegaloviruses: from molecular pathogenesis to intervention*, vol 1, in press. Caister Academic Press, Norwich, United Kingdom.
 62. Alcendor DJ, Zong J, Dolan A, Gatherer D, Davison AJ, Hayward GS. 2009. Patterns of divergence in the vCXCL and vGPCR gene clusters in primate cytomegalovirus genomes. *Virology* 395:21–32.